UCRL-JC-125616 Abs

Task Force on Network Storage Architecture: Storage for ASCI

Kim Minuzzo and Dave Wiltzius
Lawrence Livermore National Lab (LLNL), Livermore CA
minuzzo1@llnl.gov

Position Statement

The Accelerated Strategic Computing Initiative (ASCI) is a critical element of the Department of Energy’s (DOE) response to the recent decision to end nuclear testing. ASCI will provide computational capabilities for stockpile stewardship without nuclear testing by pushing the computing industry to develop high-performance computers with processing speeds and memory capacities 1000's times greater than currently available. I/O is a critical problem because of increasing imbalances between processing and memory capabilities, and I/O and storage devices. While performance of commodity microprocessors is improving at rates in the range of 50% to 100% per year, device I/O performance is only improving at a rate of 35% or less per year. Parallelism and other mechanisms must be exploited to obtain the levels of I/O performance required by ASCI.

Through ASCI, LLNL will have a system on January, 1999 with the following: 4096 PowerPC processors; 1.0 TFLOPS/s sustained (3.2 TFLOPS/s peak); 2.5 TBs of memory; 75 TBs of local disk. To maintain a system balance between this parallel processor and external I/O we predict the need for at least a 1 PB archive that can sustain a transfer rate of 8-10GB/s. In the year 2003 ASCI machines are targeted to be capable of 100 TFLOPS/s and have 50TB of memory requiring a 100-200PB storage system with a transfer rate of 100-150GB/s. To achieve these capacities and transfer rates a scalable, network-centric architecture that supports striped, parallel I/O across large numbers (100-1000’s) of network attached devices is required. The plan for developing this storage system is to mirror the development of the ASCI compute platforms: commodity components coupled together to form systems with the required performance. The network attached storage device is a key component for our planned architecture.

The DOE Labs have explored various network storage issues in a heterogeneous computer environment. Based on our experiences we offer the following observations regarding:

Data throughput. We achieved 30-60MB/s with different computers using IPI-3 over HIPPI. To obtain this performance customized API, HIPPI hardware, and IPI-3 drivers combined with a high-end RAID system were required. TCP/IP performance issues are understood with published literature analyzing bottlenecks on hosts. Our TCP/IP performance numbers have increased with faster workstations, new releases of operating systems, and new host adapter hardware and software. We measured I/O performance delivered to the application by a few SCSI attached RAID systems. We can achieve performance up to 18MB/s accessing the RAID systems using raw I/O, compared to 3MB/s with UNIX I/O.

Scalability in capacity and throughput. Early performance tests using a distributed HSM under development indicates that scalability is an achievable goal (data transfers occur across an IBM SP2 interconnect using TCP/IP):

Number of Clients & Disks	Aggregate Transfer Rate
16	112.1 MByte/s
32	174.7 MByte/s
64	334.0 MByte/s
128	636.9 MByte/s

240 | 1353.3 MByte/s

Data abstraction. Current I/O protocols are at the extremes of abstraction: Block I/O (e.g., SCSI, IPI-3) and file I/O (e.g., NFS, AFS). High performance I/O often uses techniques that require the file directory services be separate from the data storage server, making file I/O storage systems undesirable.

Hierarchical storage. Network attached storage efforts should be extensible to other storage media, particularly tape.

In summary:

- We consider competitively priced network attached storage strategically important to meet our scalable (capacity and performance) storage needs.
- Many bottlenecks must be resolved to utilize high speed I/O media.
- The level of data abstraction can be elevated considerably from block I/O, and be beneficial to high performance HSMs
- Anticipate network attached storage throughout the storage hierarchy.
- HSMs can utilize this scalable storage!

The LLNL work was performed under the auspices of the U.S. Dept. of Energy under contract no. W-7405-Eng-48. UCRL-JC-125616 Abs